

ASHESI UNIVERSITY COLLEGE

DATA MINING AND PUBLIC HEALTH SURVEILLANCE

BY

ADETUWO, ADEYEMI TEMITOPE

Dissertation submitted to the Department of Computer Science

Ashesi University College

In partial fulfilment of Science degree in Computer Science

APRIL 2013

DECLARATION

I hereby declare that this dissertation is the result of my own original work and that no part of it has been presented for another degree in this University or elsewhere.

Candidate's Signature:

Candidate's Name:

Date: 10th April, 2013

I hereby declare that the preparation and presentation of the thesis were supervised in accordance with the guidelines on supervision of thesis laid down by Ashesi University College.

Supervisor's Signature:.....

Supervisor's Name:

Date: 10th April, 2013

Acknowledgments

Sincere gratefulness goes to my supervisor, Mr Aelaf Dafla, for his assistance and feedback throughout this work. I owe a lot to Mr Dafla as this research project was his idea when I was struggling to come up with what I wanted to work on. He was my source of inspiration throughout, as he was always available whenever I had issues with this project. I truly appreciate everything he has done.

Finally, I would like to thank my family, friends and colleagues who have supported me during the course of this dissertation. Without their support this work would not have been possible.

Table of Contents

DECLARATION	1
Acknowledgements	2
Table of Contents	3
ABSTRACT	vi
CHAPTER ONE	1
1.1 Introduction and Background.....	1
1.2 Data Mining Tasks.....	2
1.3 Motivation.....	3
1.4 Related Work.....	3
1.5 Thesis.....	3
1.6 Research Question	4
1.7 Objectives.....	4
1.8 Limitations of This Study.....	5
1.9 Outline of Thesis Report	5
CHAPTER 2 LITERATURE REVIEW.....	6
2.1 Data Mining Techniques.....	6
2.1.1 Artificial Neural Networks.....	6
2.1.2 Classification	7
2.1.3 Clustering	7
2.1.4 Decision Trees	7
2.1.5 Association Rule.....	8
2.2 Limitations to Data Mining.....	8
2.3 Data Mining Issues	10
2.4 Data Mining and Public Health Surveillance.....	12
2.5 What is Medical Data?	12
2.6 Uses of Medical Data.....	13
2.6 Issues with Health Related Data Collection	14
2.7 Data-to-Knowledge Spectrum.....	15
2.8 Information Management Systems	15

2.8.1	Relational Database.....	15
2.8.2	Object-Oriented Database	16
2.8.3	Knowledge Base:	16
2.9	Knowledge-Based Systems	17
2.10	Types of Knowledge Based Systems	17
2.10.1	Expert Systems.....	17
2.10.2	Case-based Reasoning	19
2.10.3	Intelligent Agents.....	19
2.11	Real-Time Knowledge-Based Systems.....	20
2.12	Coding Systems	22
2.13	Knowledge-Based Systems in Health	23
Chapter 3	– Problem Definition	27
3.1	WEKA.....	28
3.2	The District Health Information System.....	29
3.3	DHIS 2 System Overview.....	30
3.2.1	Frameworks.....	31
3.2.3	Tools.....	32
Chapter 4	– Data Analysis and System Implementation.....	34
4.1	DHIS2 Architecture.....	34
4.2	OPENMRS.....	35
4.2.1	Installing the OpenMRS	35
4.3	OpenMRS-DHIS Integration	35
4.4	Data Collection and Pre-processing	37
4.5	Data Visualization.....	37
4.5	Experimentation.....	39
Chapter 5	– Discussion and Conclusion	43
5.2	Challenges.....	44
5.3	Conclusion and Future Works.....	46
Works Cited	49

ABSTRACT

The ability to observe current trends in health information and predict the future health status of a country is an important goal of health service institutions. Health information systems are used to accumulate, validate and examine data in order to provide information which is used in the formation of health policies. The fast growth and integration of databases means health scientists now have a huge resource that can be studied to make inferences and reveal valuable trends.

Developing countries like Ghana often lack such systems, with high cost of software and IT infrastructure being the major issues. However, due to its low hardware requirements, an open-source system like DHIS2 solves these problems.

There is also the need to develop systems which simulate the human thought process and data mining is one element of this stimulating area of adaptive behaviour and machine learning.

The main research objective of this thesis is to examine how medical data are being used currently and explore better ways of analyzing them to deliver an improved quality of health service.

Key concepts: Data Mining, Health data, Knowledge-based system.

CHAPTER ONE

1.1 Introduction and Background

Data mining is the science of extracting meaningful information and relationships from large sets of data. These relationships are usually termed models or patterns, and examples include graphs, linear equations and clusters (Hand, Mannila, & Smyth, 2001). It is a new subject and has its branches in data management, databases, artificial intelligence, machine learning and statistics, all of which have to do analysis of data. Over the past few years, there has been a huge increase in available digital data and consequently an increase in the size of databases. It is no surprise then that there has been interest in monitoring and analyzing these data to find out patterns that will be of benefit to the owner of the database.

It is widely known as an important tool employed by modern business because of its ability to convert data into business intelligence. This gives an informational edge. It has a number of other uses though, like surveillance, marketing, and fraud detection. The process of finding these relationships and reaching a meaningful conclusion requires a number of procedures:

- Knowing the kind of representation required.
- Finding groups in the data that are related.
- Looking for relationships between variables.
- Determining how to measure how appropriate the representations employed are.
- Picking an algorithm to find an optimal score function.
- Finding out how to implement the algorithm efficiently.

1.2 Data Mining Tasks

It is only fitting that we consider the tasks involved in a data mining activity and put them into categories.

Exploratory Data Analysis: The aim here is to explore the data without any constraints as to what the outcome should be. Explorative data analysis places a lot of emphasis on visual feedback and interaction. Pie charts and graphs are examples of explorative data analysis.

Predictive Modelling: This has to do with building a model that uses the known values of some variables to predict the value of another variable.

Descriptive Modelling: This is an attempt to describe all the data. A descriptive model is a summary of the most important parts of the data irrespective of the size of the data set. Descriptive modelling usually involves segmentation and cluster analysis.

Retrieval by Content: This refers to a situation where the user has a desired pattern, and hopes to find a similar pattern in the available data set. For example, the user may have an image in mind and try to find another image from the data set that matches it.

Detecting Patterns: This is an aspect of data mining concerned with discovering patterns. One use of pattern detection is in astronomy where the identification of unusual stars plays a huge part in the discovery of previously unknown phenomena (Hand, Mannila, & Smyth, 2001).

1.3 Motivation

The ability of machines to work together and with humans to generate valuable result based on structured data is quite intriguing. The possibility of relating the data mined in one community with the medical cases encountered in another, and solving such cases was another reason I decided to undertake this project. Because of time constraints and limited programming efforts, it is necessary compare the outcome of this study to logical human thought. It will be particularly interesting to make conclusive inferences from data sets. It is almost impossible to cover all human errors, seeing as people react differently when confronted with different copies of the same data. However, the use of a natural language processing method, enabling computers to make meaning of human input, will only help to achieve a more truthful picture.

1.4 Related Work

Over the past decade, quite a number of significant methods and solutions have been proffered. The effectiveness of these solutions is determined by how cost effective they are. There was an attempt to explore the implementation of semantic web technology in the health domain. Also, two colleagues of mine have decided to develop a mobile nurse for their project.

1.5 Thesis

My thesis is as follows:

Given the idea of routine data, there is the possibility of mining the data and integrating a health information system with a hospital system to improve the quality of information available on already existing epidemics and prevent looming ones.

1.6 Research Question

The research question for this study was borne out of the desire to understand how data mining works and how it can be used to improve the quality of health services in the country. The research question: *"Using existing data collected routinely, how do we compute, mine and predict health indicators like epidemics?"* seeks to provide a good understanding of the concept of data mining and look at the possibility of using medical data to make informed decisions.

1.7 Objectives

The aim of this work is to probe the thesis above and try to find out which approach is best to health surveillance. I am however bounded by the health standards and regulations stipulated by the Ghana Health Service, Ministry of Health and the World Health Organization.

As such it is important for me to obtain my data from professional medical sources. I have decided to limit the scope of my work to the following:

1. Assess resource organization methods and make a recommendation that highlights which is best for effective health information management.
2. Explore the use of data mining tools to help improve health surveillance.
3. Examine the possibility of integrating a hospital records system with Information Health Software to effectively gather and usefully evaluate clinical data.

1.8 Limitations of This Study

This paper uses data which relates only to Ghana and as such no assurance that the inferences made at the end will be pertinent to another country. Furthermore the size of the data set might not be as large as necessary for some of the data mining techniques and so the conclusions may not be totally spot on.

1.9 Outline of Thesis Report

The next chapter, chapter 2 consists of the literature review of the health surveillance and knowledge based systems. Chapter 3 will discuss the methodology. Here I will talk about what kind of study this work is and what kind of data will help me answer the questions. The concepts surrounding this work will be explained in more detailed terms with relation to Health Surveillance. Chapter 3 will also consist of the procedures to be followed and what tools will be employed for data processing. Chapter 4 contains the experimentation and implementation phase and the final chapter, 5 is for discussion and the conclusions I drew from experimentation.

CHAPTER 2 LITERATURE REVIEW

2.1 Data Mining Techniques

There are a number of techniques involved in the identification of patterns in huge databases.

2.1.1 Artificial Neural Networks

An artificial neural network, ANN is in the form of the human neural system and tries to copy the way human beings make decisions. Artificial neural networks have a number of advantages – the networks can derive non-linear relations and a function need not be defined beforehand to match the data. An ANN uses previous knowledge to draw conclusions and a large collection of data is not usually required. However, in cases where data is missing the ANN can still learn and make accurate predictions (Adefowoju & Osofisan, 2004).

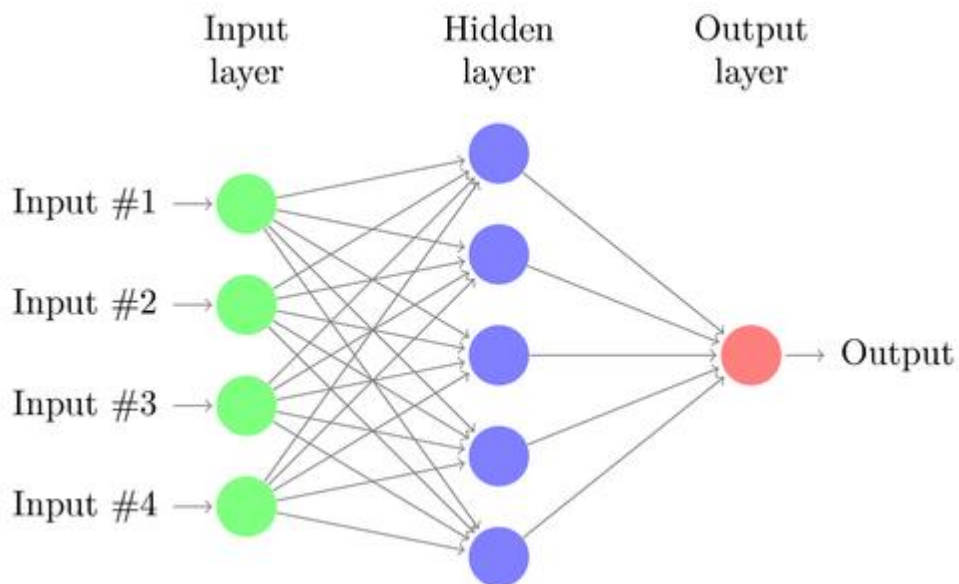


Figure 1 – An illustration of a neural network.

An ANN consists of x layers of neurons, including the input and output layers. The input layer, the first layer receives and sends external signals; and the

output layer transmits the result of the calculations. The hidden layers help an ANN to deal with difficult problems.

2.1.2 Classification

This refers to the process of allocating items in an assembly to into groups. In classification the classes are known already and the algorithm attempts to predict the class of each record in the collection. Classification of items is done in order to be able to predict for each record in the collection what the target class is. An example is a classification model that sorts insurance policies into low risk, medium risk and high risk (Oracle, 2008). For a model that is set up to predict health insurance risk over a period of time, the data will trace health history, occupation, years of residence, and so on.

2.1.3 Clustering

A cluster refers to a grouping of a number of similar things. The items in the group are referred to as members and the centre of the group is called a centroid. Clustering is done so as to divide the collection into smaller groups of similar data items. Clustering has a number of applications, including market research, pattern recognition, and customer profiling (Anderson, Benjamin & Fuss, 1997).

2.1.4 Decision Trees

This is a data mining technique which uses a tree-like model of decisions based on conditional probabilities. Decision trees are used to make guidelines and a typical guideline is a conditional statement that is easily applicable in a database and is comprehended by humans. A decision tree comprises three types of nodes – decision nodes denoted by squares, end nodes denoted by triangles and chance nodes denoted by circles. Decision trees are usually used in management

science in order to find out the best course of action to take in achieving a goal. They are also used of describing how to compute conditional probabilities.

2.1.5 Association Rule

This has to do with finding out important relationships among the items in a collection. These relationships are expressed in the form that the presence of a certain data item indicates, to a certain degree of likelihood the presence of another data item. This technique is common in text mining and bioinformatics amongst other fields (Anderson, Benjamin & Fuss, 1997).

Other tools that may be employed in data mining include:

Nearest Neighbour – This method categorizes a data element based on the data elements that share the highest similarity with it in a database.

Rule Induction – Using statistical relevance to mine valuable if-then rules from the data set (Suherman, 2010).

Genetic Algorithms – This refers to techniques of optimization which have mutation, genetic combination, and natural selection as their underlying concepts.

2.2 Limitations to Data Mining

The data mining tools discussed in this chapter are without doubt very powerful, but they have suitable conditions in which they function well. In order to be fully effective data mining requires skilled personnel for the technical and analysis part. The personnel will be able to interpret the result obtained and construct the analysis appropriately. Most of the problems that limit the efficiency of data mining are connected to unavailable or insufficient data, absence of skilled personnel and ineffective technology.

On the conditions required for each tool to work effectively, decision trees are able to detect patterns clearly so they can be employed when pattern detection is the objective. Cluster algorithm can be used when trying to identify groups within a data set that share similar characteristics. Neural networks are mostly used when predicting patterns in data sets. For all its ability to predict and detect relationships, it does not tell what the patterns mean or how they can be used meaningfully. Also the strength of these relationships hinges on how the real world operates.

In order to explain this, we consider a data mining application aimed at identifying potential burglars in a group of people. In order to get an accurate result, the model used may employ already existing data about known burglars and their characteristics. Even with this approach, it is possible for the model to miss a burglar who exhibits significantly behaviour.

Another limiting factor to data mining is that it does not show a causal relationship to the pattern(s) it detects earlier (Seifert, 2004). In order to explain this we consider a popular behaviour, buying airline tickets just before the plane is scheduled to take off. It is natural to attribute this behaviour to one of three factors – level of literacy, level of income, frequent use of the internet. That these three factors are most likely does not make them the only reasons why people book last-minute tickets. When we consider that certain people like to take advantage of last-minute discounts for the fun of it, then our data set becomes bigger. It is also possible that the person has an emergency they must attend to at that point in time and that explains why they booked a late ticket (Potomac, 1994).

In view of these limitations, in order to derive sensible and useful patterns, the researcher has to carry out a critical analysis of the results from the experiments.

2.3 Data Mining Issues

There are a number of issues associated with data mining and these issues are connected to lapses and execution. They include, but are not in any way limited to, mission creep, privacy, quality of data, and interoperability (Seifert, 2004).

Mission Creep: This highlights how fragile a suggestion control over one's information can be. Mission creep refers to the use of data for other purposes different from that for which they were collected in the first place. It can occur whether or not the individual provided the data by their own will. For example, an effort to combat terrorism can become so pressing that the officials in charge of the data are put under a lot of pressure. They will be forced into using every data resource available at their disposal, so as not to appear careless. They will feel the need to make available any data that may lead to the end of a current threat or future threats.

Government officials who are charged with national security can be made to compare or combine databases to detect any threats. Usually, searching for and accessing information for other purposes than originally intended is inoffensive; but the information from used from this search sometimes give false results. The main reason given for this is incorrect data (Seifert, 2004).

Every data collection effort is susceptible to varying degrees of inaccuracy and it is quite expensive to ensure the accuracy of the data. In controlled environments, the organization collecting the data is well aware of the limitations of the data being collected. However when the same data is taken out

of this controlled environment and employed to perform another task, irregularities occur and false inferences are made (Seifert, 2004).

Privacy: With the advent of information sharing techniques and data mining enterprises, there has been a huge attention towards the impact of these on privacy. Privacy concerns are noticed on original projects planned for and for cases of mission creep. An example of a case where privacy issues crop up is in counter terrorism agencies. Some trade-offs are expected, regarding whether or not to sacrifice national security for privacy. Even though existing laws do not suggest that data mining poses any threat to privacy, there is not enough information on how the projects are undertaken and more supervision is required. Clearer rules and better supervision of data mining efforts will go a long way in helping to solving privacy issues.

Quality of Data: This refers to the degree of correctness and wholeness of data. It is a multifarious issue and it is influenced by the organization and regularity of data being examined. Subtle differences in data have a huge effect in complex data mining techniques; and these differences are brought about by human error, duplication of records, and poor updating. In order to improve data quality 'cleaning' the data is required. This involves eliminating duplicate records and removing redundant data fields from the database.

Interoperability: This refers to the capability of data to work with data in other systems under common standards. Interoperability is geared towards encouraging collaboration between agencies and fostering information sharing through government projects. Interoperability is required in data mining because searches usually involve combining and comparing multiple databases.

2.4 Data Mining and Public Health Surveillance

Public health surveillance refers to the continuous and systematic gathering, analysis and interpretation of medical data which is then sent to health institutions. The information derived from surveillance makes it possible to give quick responses to the health needs of the community. Public health surveillance provides early signs to the institutions in charge of preventing and controlling diseases. It is important in preventing disease outbreaks and gives the institutions ample time to plan before it becomes an epidemic. Surveillance is geared towards making accurate health calls and it is necessary to develop a system to help the process. Such a system should take data derived from laboratories and clinics and provide early warning.

2.5 What is Medical Data?

The idea of an individual being medically ill can be closely linked to the accumulation, examination and understanding of their vital statistics. In early Greek literature, collecting and interpreting these statistics are central to the process of health care (Shortliffe & Barnett, 2006). This is because these statistics are crucial to the decision-making process and they provide the grounds for grouping the problems faced by a patient. Shortliffe and Barnett describe a medical datum as any observation about a patient, for example the blood pressure.

Data, being plural simply refers to multiple observations of this nature. A datum generally has four elements – the patient, the time of the observation, the parameter being observed (blood pressure for example) and the value of the

parameter (Shortliffe& Barnett, 2006). Medical data can come in the form of textual data, numerical measurements, drawings, and sometimes photos.

2.6 Uses of Medical Data

Public health communities gather data in many ways depending on the type of health event being discussed. The most popular one is that of notifiable diseases, a system based on people reporting diseases to the health department for proper treatment and control (Begg& Berlin, 1988). Another system used to obtain data is from hospitals and their extraction techniques. Because data are usually collected for other purposes, they must be adapted before they can be used in surveillance. According to Colin Begg and Jesse Berlin, statistics have nearly complete level of certainty but they cannot be used for surveillance purposes due to inadequate information. Certain data, like occupation and salary might not be present.

Data for surveillance can be gathered from valid on-going surveys and whatever method of collection is employed, it should be systematic and suitable for its purposes. Health data is analysed for a number of reasons:

- To estimate the magnitude of a problem.
- To determine the geographical distribution of a certain disease.
- For early detection of epidemics.
- To evaluate the effectiveness of control and preventive measures.
- Keeping a tab on changes in infectious agents.
- Identify deviations from expected trends.
- Support clinical research.

2.6 Issues with Health Related Data Collection

Information about health consumers is very important to health workers so as to ensure their best that the clients get the best form of treatment.

With the availability of internet connection, patient health information can be easily retrieved. However, gathering information with technology has its problems; problems which do not arise in the traditional data collection approach. I shall examine the concept of an integrated health record system and the problems encountered by worker in gathering health data.

An integrated health record system is a form of database which takes all health records of health consumers in a location within a given time period (Quynh, 2006).

This system is needed for a number of reasons: there is the need for health workers to access similar information so they have enough information about their clients. These clients do not always know their health conditions and more often than not, they do not remember their health history. Also testing outcomes derived at different health centres is possible if there is a common system where they put all their information. One last reason is that an integrated health record system is of great interest to research and health management.

Quynh Le believes that if there is a record system, policy formation is much more informed, interventions yield better results and are less expensive. It is also much easier to identify the causes and risks associated with a disease, and to monitor disease outbreaks. There is an established inventory of disease, causes, and how it can be treated.

Setting up a health record system is not an easy task because there are some issues encountered in its implementation. The information must be truly

available and usable at the point of need. But it is not enough for information to be available when needed, it is very important for the information to be classified and protected and guided by the wishes of the client. It is very important for the system to be frequently updated in order to keep up with changes in patient information.

The problem of updating arises when there is the need for a mechanism to monitor the process of updating individual records. Running a health record database is a very expensive project, because it requires the use of servers and networking on a very large scale. Another problem is security; unauthorized users of the internet and intranet can intrude the network and monitor network traffic, having access to information they should not.

2.7 Data-to-Knowledge Spectrum

A huge focus in medical research is a form of information base that explains virtually all medical terms. Medical personnel and researchers are working on showing the differences among medical data, medical knowledge and medical information (Shortliffe & Barnett, 2006). Medical data is simply observational points taken from a patient at a particular time. Analysis of this data with the aid of models, assumptions, facts and rules leads to a form of knowledge.

Information is simply organized data ready to be analysed.

2.8 Information Management Systems

2.8.1 Relational Database

A relational database refers a collection of data items organized as a set of formally-described tables from which data can be taken or rearranged in a number of different ways without the need to reorganize the tables in database. A relational database is a set of tables that contains data built into categories

defined initially. Every table in the database has one or more categories of data in columns.

2.8.2 Object-Oriented Database

An object-oriented database management system is a management system in which data can be modelled and created as objects. It also has classes of objects and class properties can be inherited by subclasses. An object oriented database should normally make use of object-oriented programming languages like Java and C++. The features of an object oriented database include persistence, concurrence and recovery.

Also, management of secondary storage, complex objects, object identity, encapsulation and inheritance must be possible in an object oriented database.

2.8.3 Knowledge Base:

A knowledge base basically serves as a storage facility for information. Here, information is gathered, arranged, distributed, searched and used. A knowledge base can either be set up to be used by machines or for human use (O'Leary, 1998). Machine-readable knowledge bases save data in a format that makes it easy for a computer to read. This is done to employ the machine's help in arriving at a logical conclusion. A human-readable knowledge base is structured in a way that allows people to access and use the information contained in them.

There is a difference between a database and a knowledge base; a database simply stores data without any form of concluding analysis. Meanwhile, a knowledge base describes a collection of facts, rules, models, that can be applied in solving problems and analyzing data (O'Leary, 1998). It is possible for a knowledge base to provide enough information and structure which includes a

semantic relationship among its fields. In such cases, the system can use this knowledge and relationships to solve medical cases.

2.9 Knowledge-Based Systems

Knowledge-Based Systems, KBS are tools which take their framework from methods of artificial intelligence. They work in constricted fields to help in decision making process. A knowledge-based system enables a better informed decision because its knowledge is documented, it learns by itself and it gives justification for each suggestion it makes. The basic components of a knowledge-based system are – a knowledge base, an inference mechanism and an acquisition mechanism.

2.10 Types of Knowledge Based Systems

2.10.1 Expert Systems

Expert systems convert human knowledge to code. Imagine an application that the sort of service one would usually get from a counsellor. A counsellor basically helps with decision making, using available information to make his or her work easier. Expert systems, a section of the much bigger field of Artificial Intelligence perform a similar role.

These systems exhibit human-like characteristics and they make use of people's experiences. Their range of operation is still quite narrow at the moment. An expert system usually consists of a working memory, a user interface, a knowledge base, an explanation system and an inference engine (Anjaneyulu, 1998).

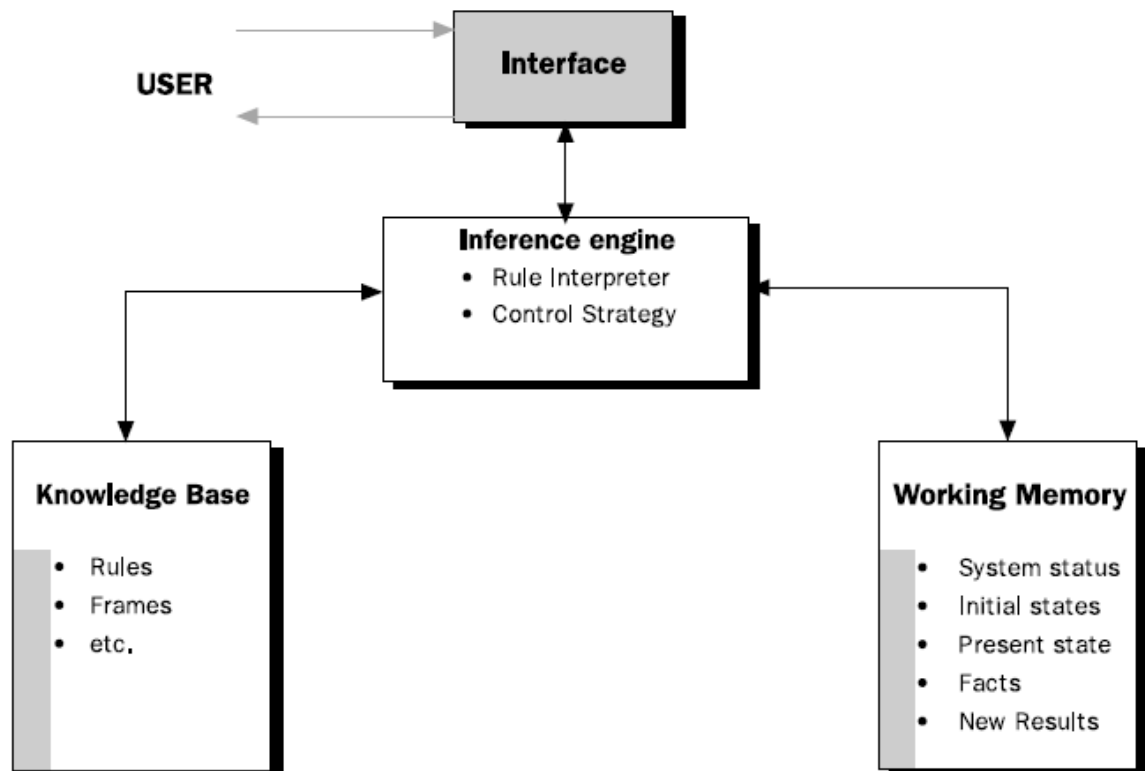


Figure 2 - Components of an Expert System.

The working memory refers to the collection of information within the domain; every fact associated with the domain (human knowledge and experience) is stored in the working memory. The inference engine is the part which contains the program on which the system runs. It uses the relationships in the knowledge base and the experiences in the working memory for problem solving.

The user interface provides an avenue for the user to supply data to the system and for the system to pose questions and offer solutions. An Expert system has to be able to provide a logical basis for its conclusions; hence an Explanation System. Since the system uses human knowledge and questions to make its decisions, the user should be able to ask why a particular question is necessary and how the system arrives at a particular conclusion. This makes it easy for the human user to test whether or not the reasoning of the system is spot on.

2.10.2 Case-based Reasoning

Case-based reasoning, CBR refers to a methodology applied in building smart computer systems in which a new problem is solved using solutions applied to an older problem (Riesbeck & Schank, 1989). In CBR old cases are saved in memory and are retrieved when a similar situation comes up. The experience is reused in the context of a new case, with the level of reuse dependent on the similarity between the cases.

A typical case illustrates a situation of diagnosis and is made up of a description of the symptoms, the cause of the failure, and how the fault can be repaired (Riesbeck & Schank, 1989). Take a car fault for example; one of the symptoms observed may be the refusal of the engine to start. The next course of action is to find the cause of the fault, which might be an empty car battery in this case. The fault can then be repaired by getting a new car battery or charging the empty one. If the diagnosis made is correct, it is stored in memory with a view to applying it to a similar case in future.

2.10.3 Intelligent Agents

Intelligent agents are units, computer programs that carry out useful tasks on behalf of humans, displaying some unique properties. An intelligent agent must understand its purpose, interact with any environment in which it finds itself and react to the smallest change in that environment. An intelligent must exhibit degrees of intelligence, a capacity to make decisions and learn from incidents (Zane, 1999).

Some Intelligent agents help in collating information and reporting the outcome of this action. Others, like robots carry out tasks that affect the world directly.

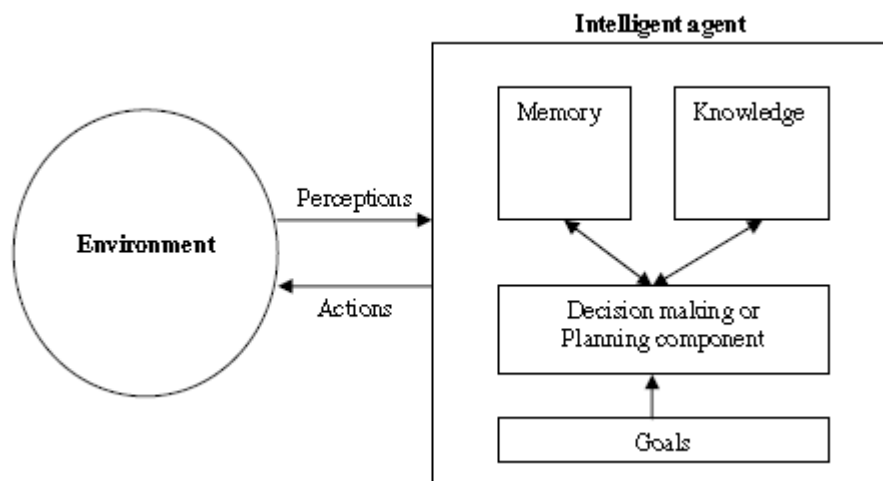


Figure 3 - General architecture of an intelligent agent (Zane, 1999).

2.11 Real-Time Knowledge-Based Systems

Real time presents a new interesting platform for applying intelligent problem solving methods. This section examines how problems encountered in the Real-time domain are different from those experienced in other fields. Many definitions of “real-time” exist, but perhaps the most popular use of the word implies being “fast”. This means that a system is said to be real-time if it processes data rapidly (O'Reilly and Cromarty, 1985). However a system can be considered to exhibit real-time behaviour if it is “predictably fast enough for use by the process being serviced” (Marsh & Greenwood, 1986).

Real-time domain poses the toughest situations and a knowledge-based system operating in real time (for example a crisis control mechanism) must be able to respond immediately to changes in its environment. It should be able to do this while considering the resources available to it (Laffey, Cox, Schmidt, et al, 1988). Resources include hardware capability, available memory, and time allotted amongst others.

The major advantage of using a real time system is to minimize the knowledge load on users or to allow them boost output without increasing the knowledge

load on them (Turner, 1986). In the presence of a real time system, human beings do not need to memorize parameters and they are not required to relate figures on the go. This implies a real time system will be of good use when an individual cannot capably monitor every data available, when the individual is not capable of avoiding or solving clashes, when the individual cannot multitask proficiently, and when the response time is too high. An example of a situation when this occurs is in a Stock Market where humans are confronted with a lot of information and figures, and they have to make accurate decisions on the spot. For a system to be considered real-time, it should meet some requirements. It should be able to focus should a major event occur and its behaviour should be predictable. It should make best use of the surroundings it operates within, and the system should be able to continue working even if there is a fault with one of its parts.

As interesting as real-time tools seem, there are limitations to the way they operate. The tools have some complex and unstable problems because of their huge dependence on time. Real-time tools have no ability to handle interrupts between software and hardware. The possibility should exist for efficient management of asynchronous messages.

This means that an outgoing message can be interjected and continued when another message with a higher priority is being processed (Laffey, Cox, Schmidt et al, 1988). They cannot be incorporated with a real-time clock, and they can only receive input from human stimulus. The tools are on hardware which is not built to resist harsh conditions and make break down in the face of such.

A lot of work is being done to develop real-time systems that are intelligent. X-NET, a network-management system intended to be used with complex data network networks. These networks usually cover a large area and are expected

to work 24 hours. X-NET is in charge of observing and analyzing the efficiency of the network by checking network activity and error rates amongst other things. It diagnoses data sent from checking network activity and records any diagnosis made. Diagnoses are kept in logs until the operator stops the process. Real-time systems help solve the problem of human limitation, which is arguably the biggest reason why we need them.

2.12 Coding Systems

Edward Shortliffe and Octo Barnett, in their paper titled: "Medical Data: Their Acquisition, Storage and Use" suggest that medical personnel have used routinely collected data like absenteeism, emergency room visits, and over-the-counter purchases to monitor and provide early warning of an epidemic. But the paper goes on to state that the use of this approach will depend hugely on considering how epidemics affect such data.

The development of computerized methods of detecting epidemics has led to new methods of public health surveillance done by analyzing routinely collected data (Shortliffe & Barnett, 2006). This new model is based on the supposition that the effects will show early and the data obtained from this event will be incorporated and analysed rapidly and effectively.

This approach requires a proper understanding of the effects of epidemics on the data. For example the algorithms to be employed in this early detection will depend on the characteristics of the data. Analysis of routinely collected data will be helpful in detecting rare diseases very early. Because diseases of this nature have not occurred in a long while, there is little knowledge on how they affect absenteeism, emergency room visits, and over-the-counter purchases.

Another form of reporting involves coding all hospital activities, like the kind of surgeries performed, and the kind of diagnoses made. The codes are reported to the health agencies and may also be used by the hospital.

2.13 Knowledge-Based Systems in Health

The vision of developing cutting-edge computer systems that would imitate human reasoning has led to the introduction of knowledge-based systems in medicine. This part of my work looks at the role of KBS in medicine thus far and tries to predict the trend in the future. There has been a huge rise in the interest around medical KBS over the past decade and this is because they offer a viable alternative to solve problems that cannot be dealt with using conventional techniques (Metaxiotis & Samouilidis, 2000).

Medical personnel will react differently when confronted with different copies of a patient's medical information (Kalogeropoulos, Carson & Collinson, 2002). This difference in inference has consequences ranging from increased cost of medical service to inability to properly satisfy the patient. As such there is the need to stimulate a homogenous decision making process from all medical personnel that view the same medical information.

There are a number of advantages to this process; the quality of healthcare improves, it saves cost and it guarantees that both the patient and the public are satisfied. Simply put, homogeneity in decision making leads to a cost-effective and dependable medical practice.

One step towards this is Electronic libraries, but sadly the information derived from these libraries has not been employed in any form of online medical tool. This is because a system that will recognize medical information and medical knowledge as medical objects has not been developed. Dimitris Kalogeropoulos,

Ewart Carson & Paul Collinson, in the paper discuss the development of an intelligent medical information medical system. In the most basic form of medical decision, the system should be able to collect data in an organized manner and provide a platform for experimentation for developing and evaluating the decisions made by knowledge-based systems.

KBS usually use representational reasoning rather than just calculations, and they have a robust knowledge base which features keywords that make sense to a professional in that field. KBS are usually able to support their conclusions with an explanation that makes sense to the user.

Artificial Intelligence in medicine is mainly concerned with the creation of artificial intelligence systems that carry out diagnoses and make recommendations on therapies (Clancy & Shortliffe, 1984). A medical programme is usually in phases, diagnosis, therapy and rehabilitation. Sometimes prevention of the disease is included as one of the phases.

Medical KBS help to generate reminders; a knowledge-based system synchronized with a monitor helps to transmit signals concerning a patient's state of health to a medical personnel. KBS can also scan and examine patients result from a laboratory procedure.

In situations when a patient has a rare condition and diagnosis is proving difficult to make, a knowledge-based system can step in by inferring the most likely diagnoses using the data provided by the patient. KBS can also find discrepancies in a patient's treatment schedule and they are also used to train medical students on various tasks.

An example of a medical knowledge-based system is ONCOCIN, a clinical decision support system invented in 1979. ONCOCIN is the result of an attempt to increase the ability of already-existing systems (Langlot & Shortliffe, 1983). It employs artificial intelligence techniques to help doctors with prescription and patient testing. It was developed to be used following diagnosis in order to better manage cancer patients receiving chemotherapy.

The steps and procedures involved in managing a cancer patient become significantly complex over time and it is quite tasking to memorize them. ONCOCIN takes in progressive records of a patient's treatment and compares it with its knowledge base, containing protocols for prescription and testing. This feature makes decision-making easier for doctors when it comes to managing specific patients.

Other famous medical knowledge-based systems include:

MYCIN: This was a knowledge-based system developed in the early 1970s to identify infections caused by bacteria, like meningitis and cholera, and for diseases that may arise from blood clotting. The system recommended antibiotics to patients by taking their body weight into account when determining the dosage. Like every knowledge-based system, Mycin had an inference engine and a simple knowledge base. The physician answered a series of yes/no questions, and from the responses the system showed the possible bacteria and diagnoses that can be inferred. It also provided its confidence in and a logical explanation for each diagnosis. Due to certain ethical and legal issues though, the system was never used in practice (Heckerman & Shortliffe, 1992).

DXPLAIN: This is an online medical support system that helps medical personnel make diagnoses. The system takes the symptoms of the patient, the result from the laboratory and other discoveries made by the medical personnel. It also serves the purpose of a reference because it has a database of all diseases and symptoms recorded. As with every other KBS, DXplain offers evidence to support every diagnosis it makes. It uses a pseudo-probabilistic algorithm to produce its diagnoses in strata (Detmer & Shortliffe, 1997). Using the information stored in its database concerning the commonness of each disease, DXplain is able to tell which diseases are common and which ones are rare.

DOSECHECKER: Dosechecker was developed to help pharmacists monitor orders for drugs that must be carefully administered to patients with kidney problems. Some drugs need to be taken in specific doses, and in patients with kidney issues, drug concentrations can reach high and dangerous levels. The drugs are aimed at regulating the concentration, and Dosechecker uses patient data like patient body mass and serum creatinine level. The system checks if the values for these two indicators fall within established range and gives an alert if they do not (Jick, 1977).

Chapter 3 – Problem Definition

There are two questions to be examined in this paper –

1. *Using a mining tool, how do we use available data to get a picture of the health status of a very large community, say an entire country?*
2. *Given the idea of routine data, is it possible to integrate a health information system with a hospital system to improve the quality of information available on already existing epidemics and prevent looming ones?*

The first question involves mining health data of patients with data mining software. We need data of from people that have visited hospitals in different health districts across the country over six months. Data to be taken on each patient include:

- Age
- Nationality
- Gender
- Patient ID Number
- Location of the hospital (Town A/Town T/Town K)
- Number of visits to the hospital within the period.
- Marital Status (Single/Married)
- Allergies (Yes/No).
- Employment Status (Unemployed/Employed)
- Doctor's Diagnosis (Malaria/Cholera/Tuberculosis)
- Medical Insurance Number
- Cell phone prefix (024/026/027)

- Both Parents alive? (Yes/No)
- City of residence
- Body Mass
- Height
- Body Mass Index

The Ghana Health Service, GHS delivers an annual report which shows a summary of the previous year's operations, and a projection of expectations for the coming year. The GHS uses a health information system that provides evidence for tracking the efficiency of the Ghanaian health system and improves the quality of health services given to people. This paper intends to run the data through a data mining tool and obtain a report similar to that used by the GHS.

The purpose of this chapter is to introduce WEKA, the data mining tool to be used, and then the District Health Information System, DHIS. The chapter also takes a look at the background of the DHIS, its key achievements to date, and some of the problems involved in operating a health information system in Ghana.

3.1 WEKA

WEKA, Waikato Environment for Knowledge Analysis is a machine learning software that contains a collection of visualization tools and algorithms employed in a variety of data mining tasks. I chose WEKA because it is open source and runs on virtually any up-to-date computing platform. It also has a wide-ranging collection of techniques for data pre-processing and modelling, and its graphical user interface makes user experience pleasurable. The techniques to be used in this paper are Clustering and Artificial Neural Networks.

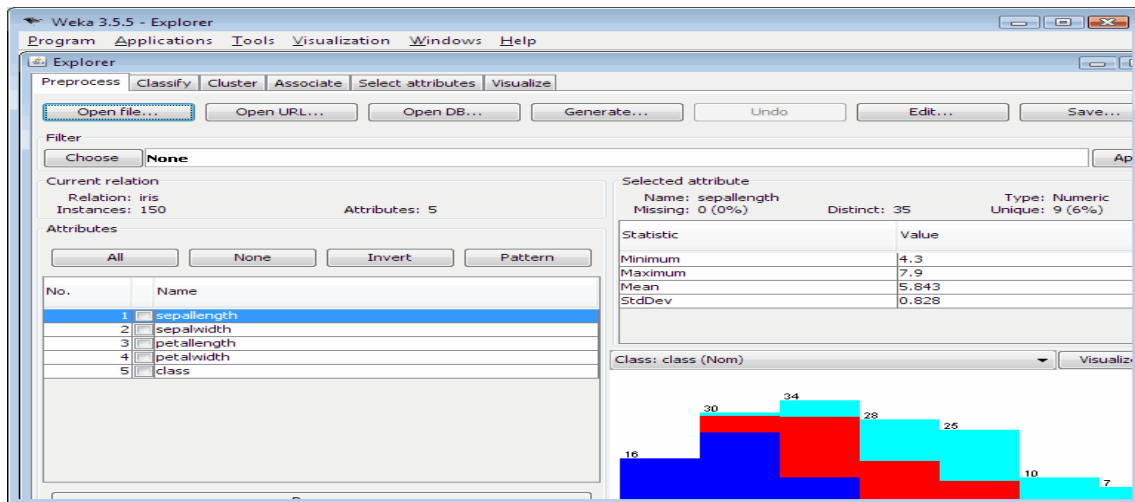


Figure 5 - Screenshot of WEKA's Explorer Window.

3.2 The District Health Information System

The District Health Information System is an open-source tool developed by the Health Information Systems Programme project. It is used for collecting, validating, analyzing, and accumulating data designed for activities involving health information management. It is a basic tool instead of the usual, a database that has been configured beforehand. The DHIS has an interface that allows the user to design the information system to taste, without having to write any code. It was originally established to cover three health districts in Cape Town, South Africa between 1998 and 1999. But its success in these districts has led to a phenomenal increase in its coverage; the HISP network now spans nearly half of sub-Saharan Africa, and some countries in Asia. Currently, the HISP network serves 300-400 million people in these regions.



Figure 5 - Screenshot of DHIS2 web-based interface.

The DHIS was set up as a medium of monitoring the delivery of health services in the public health sector. It generates data in order to help improve planning, tracking and reporting in the health sector.

The DHIS uses accumulated routine data, survey data, and some patient-based data to carry out these tasks. It has tools aimed at improving the quality of data and for data validation. It also has a dashboard specifically for each user which shows relevant tools like indicator charts used for evaluation and monitoring. The DHIS has a user management component for passwords, and users can share their data using graphs and reports.

3.3 DHIS 2 System Overview

This section examines the fundamental architecture and technologies used by the system. DHIS 2 uses quite a number of open-source software tools; an advantage of this is that it makes the system less dependent on one technology.

The system is also more flexible and combining different tools makes development quicker.

The DHIS 2 makes use of Java 6 programming environment; the choice of programming language was prompted by the desire to have a system that is compatible across different platforms. Because of the flexibility of the J2EE platform, it is used in a lot of Java applications and the DHIS 2 was built to work on any container compatible with J2EE (Steensen & Vibekk, 2006).

3.2.1 Frameworks

The DHIS 2 makes use of the Spring Framework. This provides a broad configuration and programming structure for Java-based applications. Spring framework is used to handle cases of Dependency Injection. Dependency Injection refers to a pattern that reduces the dependency coding. Rather than fully coding every unit into the application, Dependency Injection makes it possible to define the units in a configuration file, either at compile-time or at run-time (Krajca, 2010). These units are incorporated into the application using spring framework.

The DHIS 2 needs to be independent of a database and the Hibernate framework is employed to ensure this. Every essential data element has its save implementation, and this provides methods for create, read, update and delete operations. For example, the `HibernateDataElementStore` class has functions that allow adding and deleting of an element, `addDataElement` and `deleteDataElement` respectively (Øverland, 2009). Hibernate is a collection of related objects in database management systems and it is most useful in mapping database tables to Java objects and vice versa. It uses its own query

language, HQL which looks like SQL and it has been employed in relational databases like MySQL and Oracle.

JUnit, unit testing model is one other piece of software used in the DHIS 2. The development of the application has test phases and these phases are made simpler. Simpler testing encourages more testing and as such a better application. These tests are compiled and added as JAR files during compilation (Krajca, 2010).

3.2.3 Tools

Maven is a build-automation, comprehension and project management tool. It is primarily used for Java projects but it can be employed in managing projects written in languages like Scala and C#. Maven is based on the idea of a project object model, a way of building a project and the relation among its internal units. DHIS2 uses Maven to bring these units together and building a system that works (Steensen & Vibekk, 2006).

Jetty is usually used as an application servlet container for DHIS2. Jetty is a small application that can be embedded with ease in projects like Geronimo. It can be seen as a JSP/servlet holder and also a HTTP server, used when deploying J2EE applications on the web. Jetty is useful in developing servlets on a local machine. Jetty is open-source, and another software application that serves the same purpose is Tomcat (Øverland, 2009).

Another tool used in DHIS2 is Bazaar. Bazaar is a control system, used to control and direct the behaviour of other systems. It is used to make it easier for different people to work together on a project. It does this by taking records as the project goes on, providing an easy avenue to copy these records around, and making it easy for changes to be implemented between projects. Bazaar

makes undoing changes easier, and simplifies comparisons between different versions of software. It is an open source, easy-to-use and quite a flexible tool. It is developed by Canonical Ltd.

In Ghana, the DHIS2 is used at the national level and only shows a summary of the medical events for the past year, or month as the case may be. It is also used to prepare for the future and for these predictions to be accurate, the DHIS2 needs to accept data from as many hospital systems as possible. These hospital/health records systems have reports of indicators that tell whether or not there is an impending epidemic for example. This is where the gap exists in Ghana; updating the Information System that exists at the national level regularly with data from hospital systems at the grassroots.

The next chapter of this work takes a look at the methodology of the data mining process. It also touches on the architecture of the DHIS2, introduces an Electronic Health Records system, OpenMRS and explores the possibility of integrating the DHIS with the OpenMRS for improved reporting of clinical indicators.

Chapter 4 – Data Analysis and System Implementation

This chapter explains the methodology involved in carrying out the study. It also the architecture of the DHIS2, introduces the OpenMRS and explores the possibility of integrating them.

4.1 DHIS2 Architecture

The DHIS is designed to fit a 3-layered architecture, a client-server concept with the application architecture comprising data layer, application and presentation. Each of these layers should be independent on the others and this makes it possible to implement one layer in the architecture without affecting the others. The data layer helps to connect to computer data storage, usually a database server; the application layer implements functional-process logic while the presentation layer provides functionality for implementing user interfaces.

DHIS2's design comprises 42 Maven projects, 18 of which are web modules. Splitting a system into modules with limited functionality makes it easier to make changes, maintain existing modules and develop new ones (Steensen and Vibek, 2006).

The main concept for data analysis and reporting is Indicator. An Indicator is a mathematical formula made up of DataElements and numbers. It is the difference between relative and absolute numbers. For example, a scientist will not want to know only about the set of subjects his tests were carried out on (absolute) but also what these tests mean to the entire population (relative). The domain model of the DHIS2 is structured in such a way that it takes any type of data.

4.2 OPENMRS

The OpenMRS community is a network of volunteers around the world. These volunteers are from different backgrounds including finance, health care, technology and international development. They are working to build a technology, the world's biggest, and one that cuts across all platforms to support the delivery of healthcare in developing countries. OpenMRS started with a single database system in a clinic in Eldoret, Kenya in 2006 and as of March 2010, it has grown almost exponentially covering 23 countries.

4.2.1 Installing the OpenMRS

I downloaded OpenMRS Standalone 1.9 from <http://openmrs.org/download/> in March 2012. OpenMRS Standalone comes with a web server and an embedded database. It represents a great way to explore the OpenMRS providing a functionality that allows the user get a local version up and running within a few minutes. It comes with an option to download sample data to use in exploring the software. I also downloaded an installation manual and followed the instruction to properly install the software. I installed the software but it did not run at first. It complained that there was an error with the openmrs-standalone runtime file. I installed Windows 7 as a virtual machine, changed the Tomcat port in the runtime file and the software ran after the third trial of connecting. A thorough discussion of the difficulties on installing and using the OpenMRS resource for this paper is discussed in the next chapter.

4.3 OpenMRS-DHIS Integration

Integrating the OpenMRS and DHIS2 shows how reporting on medical indicators can be carried out in OpenMRS in a consistent format, one compatible with the DHIS2 which analyses the data derived from these reports on an collective level, statistically. The reason behind this is to bridge the gap between the record

systems in local hospitals and the DHIS2 systems at both the districts and at the national level.

In order to integrate both systems, four components are required –

1. An Electronic Medical Records System (like OpenMRS).
2. A system that aggregates data statistically (like DHIS2).
3. A standard system or format of reporting (in this case SDMX-HD module).
4. Mirth Connect, a system that helps integrate services using HL7 standard.

After installing the OpenMRS, the SDMX-HD module is installed on it. This module comprises two modules, a reporting module and a HTML widgets module. Each of these modules is uploaded on to the Module page under the Administration section on OpenMRS.



Figure 6 – Screenshot of the Modules Upload Functionality.

After uploading the modules, a Data Set Definition, DSD file should be created using the DSD upload functionality on the SDMX-HD module. With the patient data gathered on OpenMRS over a period of time, a DSD file can be derived.

This file should show what Indicators are being reported on and they were derived. I had a problem creating DSD file, even after adding some patient data.

For my data mining section of this thesis, the steps involved are:

4.4 Data Collection and Pre-processing

My supervisor and I had planned to visit health districts to ask for their health records but this proved to be a futile task. The health officers were reluctant to give us any data relating to their patients citing issues with privacy and that was totally understandable. As a result the data to be used in this study was generated randomly, and freely too using www.datagenerator.org.

I created fields to fit the kind of data I wanted to collect, health data collected by the OpenMRS. There were 1000 entries generated using the tool on that website and the data is what will be eventually employed in this work.

The Pre-process section on WEKA has facilities for importing data stored in a database, and data stored in either .csv (Comma Separated Values) or .arff (Attribute –Relation File Format) file formats. The entries were generated in .csv format before exporting them to WEKA. The Pre-process panel also has filters which can be used in changing the data to be mined, for example converting numeric attributes to discrete ones. It is also possible to remove attributes and instances according to a specific benchmark.

4.5 Data Visualization

WEKA provides a functionality which helps users with a graphical representation of the data to be processed. There are a number of ways in WEKA can be used in visualization. The main graphical user interface will show a histogram for the attribute distributions for each attribute selected. If the mouse hovers over the histogram different ranges and the number of samples in each range are

displayed. In order to see all the distributions at once click on the 'visualize all' button and individual colours stand for different classes. That means for the 'ParentsAlive' column, 516 people answered yes, and 484 answered no.

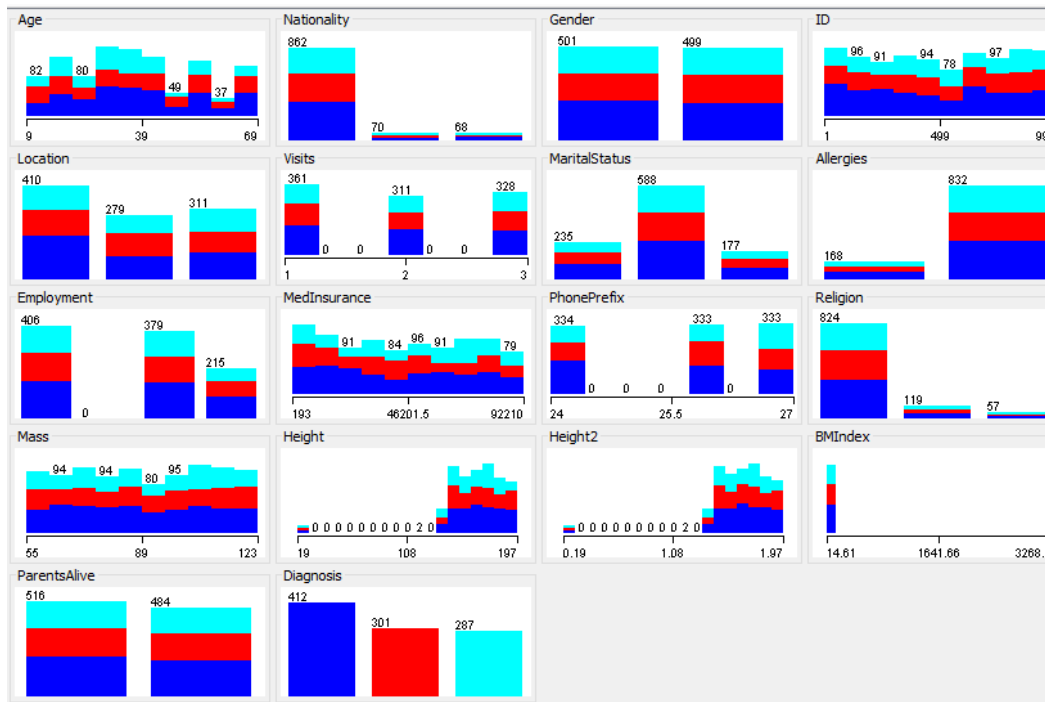


Figure 7 – Data Visualization in WEKA showing different classes.

There is also a 'visualize' tab on the WEKA Explorer page and it shows scatterplots for all attribute pairs.

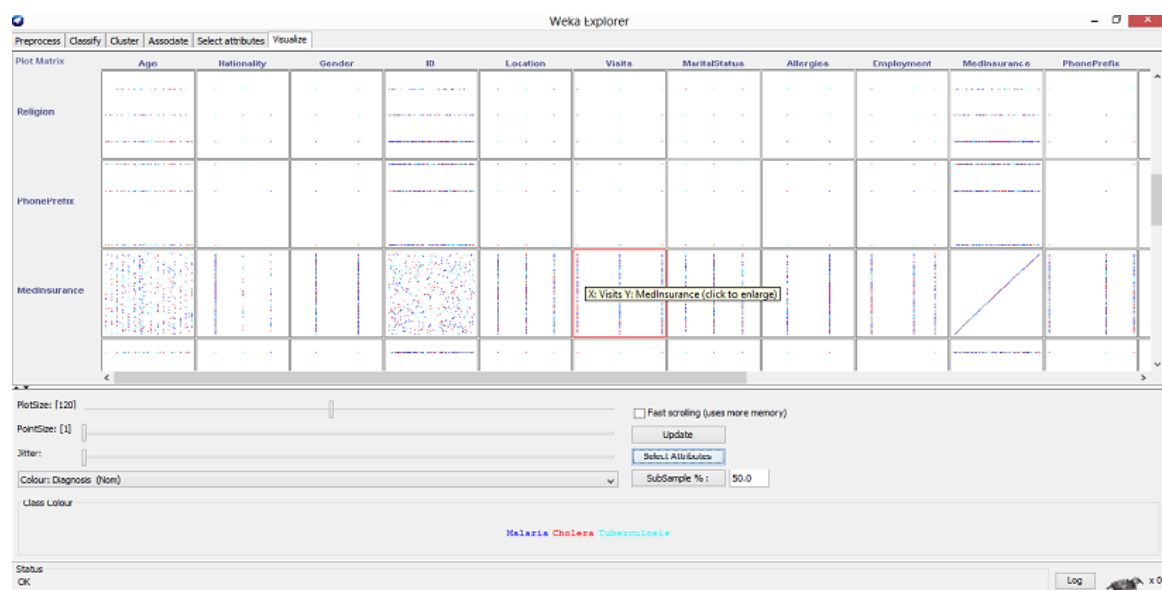


Figure 8 – Using scatterplots to visualize the dataset.

The colour is usually used together with the class attribute and colouring the other attributes helps. The colours represent different classes and in order to set your own colours for the classes, left click on the highlighted class names at the bottom of Figure 8. If we change the colour of the first attribute, age for example, we are able to discover and observe more about the data.

4.5 Experimentation

These experiments are supposed to explore the possibility of taking routine data from the OpenMRS, and using data mining software to determine the health status of the virtual country. This section of the thesis describes the results I got after implementing different algorithms in WEKA.

Artificial Neural Networks: The first technique used to mine the available health data is artificial neural networks.

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      399           39.9   %
Incorrectly Classified Instances    601           60.1   %
Kappa statistic                     0.0428
Mean absolute error                 0.4344
Root mean squared error             0.4722
Relative absolute error             99.1376 %
Root relative squared error         100.8772 %
Coverage of cases (0.95 level)     99.9   %
Mean rel. region size (0.95 level) 99.9333 %
Total Number of Instances          1000

=== Detailed Accuracy By Class ===

          TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
          0.701    0.636    0.436     0.436    0.436      0.068    0.558    0.463    Malaria
          0.163    0.162    0.302     0.302    0.302      0.001    0.479    0.287    Cholera
          0.213    0.160    0.349     0.349    0.349      0.063    0.532    0.314    Tuberculosis
Weighted Avg.  0.399    0.357    0.371     0.399    0.361      0.046    0.527    0.367

=== Confusion Matrix ===

  a  b  c  <-- classified as
289 65 58 |  a = Malaria
196 49 56 |  b = Cholera
178 48 61 |  c = Tuberculosis

```

Figure 9 – The output obtained from Logistic algorithm

The neural network used all of the health data available, with more incorrectly classified instances than correctly classified ones. The false positive, FP and the true positive, TP show the ratio of incorrectly classified instances to correctly classified instances. With an accuracy of less than 40%, we cannot draw valid conclusions using the artificial neural network. A Kappa statistic value of 0.0428 (for which 1.0 shows total agreement) is too low.

The confusion matrix has rows which show the classes in the distribution and the columns are those predicted by the network. From the matrix one can see only 289 Malaria patients, 49 Cholera patients and 61 Tuberculosis classified correctly. This indicates that an artificial neural network is not an appropriate algorithm for the health data available.

Clustering: This was the other algorithm employed during experimentation and the clustering method used is Density-based clustering.

```
Clusterer output
Normal Distribution. Mean = 25.8069 StdDev = 1.0217
Attribute: Religion
Discrete Estimator. Counts = 219 37 6 (Total = 262)
Attribute: Mass
Normal Distribution. Mean = 89.9344 StdDev = 20.6796
Attribute: Height
Normal Distribution. Mean = 162.6988 StdDev = 28.1267
Attribute: Height2
Normal Distribution. Mean = 1.6237 StdDev = 0.2803
Attribute: BMIIndex
Normal Distribution. Mean = 95.1964 StdDev = 403.0763
Attribute: ParentsAlive
Discrete Estimator. Counts = 89 172 (Total = 261)
Attribute: Diagnosis
Discrete Estimator. Counts = 64 147 51 (Total = 262)

Time taken to build model (full training data) : 0.2 seconds

=== Model and evaluation on training set ===

Clustered Instances

0      269 ( 27%)
1      311 ( 31%)
2      198 ( 20%)
3      222 ( 22%)

Log likelihood: -49.45672
```

Figure 10 – The result from Density-based clustering.

In this method of clustering, clusters are seen as the areas in the data space in which objects are dense, demarcated by areas with low density of objects. The objects in the scanty areas, needed to separate clusters are seen as border points. These areas may have a random shape, with the points inside the area being randomly distributed. This method was implemented on the health data set in WEKA and Figure 9 shows the result of the implementation. The algorithm was instructed to come up with 4 clusters and each cluster will be considered below –

Cluster 0 - This is the second largest cluster in the distribution, showing male patients from hospitals in Town T. Their average age is 35 years; they are unemployed and have been to the hospitals at least twice in the past six months. They are single subscribers on the 027 phone network, with no known allergies. They are most likely to be infected with Tuberculosis.

Cluster 1 – This represents the largest group in the distribution, depicting 34-year-old males who visited hospitals in Town A. They visited the hospitals averagely three times in the past six months. They are single, with both parents alive; and a BMI index of just over 32 suggests they are overweight. They are most likely to suffer from Malaria infection.

Cluster 2 – Another group likely to suffer from Malaria infection is the smallest group in the distribution. It represents females in their late thirties. These married, unemployed women visited hospitals in Town T once over the past six months.

Cluster 3 – The last cluster also represents females, younger women than those in the previous cluster. They have an average age of 34 years, visited hospitals

in Town K twice and are on the 026 network. They are employed, have no known allergies and are most likely to suffer from Cholera.

All the clusters give a fair indication of the health status in different regions around the virtual country over the past six months and will help the Health service institutions in planning for the future. It is important to stress that the clusters above are based on sample data and do not bear any semblance to what is obtained in reality.

Chapter 5 – Discussion and Conclusion

One of the aims of this thesis is to examine the different methods of using medical data and make a recommendation as to which is best for effective health information management. From the research I have done and the inferences I made, I think knowledge-based systems is the most appropriate for managing medical information. A knowledge-based system, like the semantic web guarantees a better informed decision because its data and knowledge are documented. It carries out machine learning, learning by itself and giving justification for each suggestion it makes.

Knowledge-based systems use data to tell what the health status of a community is, and they give reasons why they have come to such conclusions. For example, a KBS can tell the medical personnel that there is an impending outbreak of cholera in the community. It makes this assertion because of the number of cholera cases that have been dealt with in the past couple of days, and the common symptoms shared by these patients. The proximity of patients also means they may share a common water or food source and these are likely causes of the cholera.

A KBS also tells what the future will be like if the current situation is left unattended to. It is able to do this due to its experience with data over time and its ability to monitor trends.

Data mining uses quite a number of machine learning techniques but for a different reason. Typical tasks that can be carried out by a knowledge based system include interpretation, planning, design, and prediction. The solutions

proffered by the system are consistent and that makes it reliable, removing elements of uncertainty.

From the experiments I carried out, the Clustering algorithm provided the most relevant results. It gave a vivid and succinct idea of different collections of patients in all the regions. The Artificial Neural Network could not provide a specific conclusion from the dataset provided. There were a huge number of wrongly classified instances and it did not give a clear idea of what the health status is. This can be due to two reasons, the quality of data, seeing as my data was randomly generated and the size of the generated data may not be enough for both algorithms to work with.

5.2 Challenges

A knowledge based system like the semantic web has some challenges it currently faces; there is the use with developing ontologies. Ontologies are vital to the advancement of the semantic web and there has been a lot of research work focused on bringing up new ontologies. The major function of ontologies is to facilitate knowledge sharing and recycling. This is why a basic ontology library system addresses how ontologies are organized and saved to improve ease of access. Ontologies get mapped to different identifiers, and since they are expected to evolve as time goes on, it is important to ensure that different versions of ontologies are consistent.

There is the possibility of people having conflicting ideas on a subject matter and because of this validation and integrity is required. The author or source of an argument should be willing to prove that the statement is indeed credible.

The original plan was to visit districts to request for records but we could not do it. Because of privacy issues this was not done. The dummy data was generated

for the purpose of experimentation and they were a number of conflicts in the data. I believe actual patient data obtained by the health institutions will give a better representation of the health status.

The OpenMRS currently does not have enough parameters to collect information, enough fields to make mining easier. OpenMRS will be more useful for data mining if it was modified to include more information about the towns from which patient details are being recorded.

For example, there were 222 cases (around 22% of total cases) for Cholera infections in hospitals in Town K. One of the chief causes of Cholera is contaminated water and it possible that the rivers and streams in the town are responsible for this breakout. OpenMRS should have a section that relates each community with information like, what rivers/streams run through the community, availability of pipe borne water, monthly level of rainfall, etc. If such a section is available, the results generated by the data mining software will make it possible to pick out the source of the disease. Consequently this makes it easier to combat an impending epidemic.

For the next objective, I was not successful in integrating the DHIS2 and OpenMRS. I had previously installed DHIS2 and I had explored it, adding datasets and organizational units. I installed OpenMRS but it did not run at first. It complained that there was an error with the openmrs-standalone runtime file. I made every change I could but it still did not work. I sought help online and came across a comment on an OpenMRS forum from someone who experienced the same issues I had. The comment suggested that the Standalone version I downloaded is not compatible with 64-bit Windows 8 computers. I had two options, download the Enterprise version and install it on my Windows 8

computer or install Windows 7 as a virtual machine. I downloaded the Enterprise version took a while but I was able to make the software run.

The Enterprise version requires that Tomcat and MySQL be installed independently but that was not as issue because I already had XAMMP. In order to get on to the homepage of the OpenMRS this time, I was required to use the Tomcat admin dashboard. In order to do this, I needed to change the settings and include an authentication form which takes a username and a password. I edited the Tomcat configuration file that came with the Enterprise version, changing the username and password and the authentication form appeared. However upon submitting the details, the service returned an error page and did not load the Tomcat dashboard.

5.3 Conclusion and Future Works

OpenMRS-DHIS2 integration is an intended implementation and I did not come across any record of a successful implementation. That said, the theories and guidelines available suggest that it is a possibility and I think I would have done it if I had more time to work on the project. Another thing worth considering is that there is a different approach to the integration that may not present as many stumbling blocks as I faced in the execution of this work. This approach involves importing and exporting of data using excel sheets.

OpenMRS had a functionality that supports the export of data in excel format and this marks the beginning of the integration. The next step is to download the Form Data Export Module, the module that activates this functionality. This module allows the user to export medical html forms on OpenMRS to a .csv file. The design is aimed at facilitating an easy mechanism for pulling out data from OpenMRS.

If this has been done and the .csv file is ready, the next step is to import it into DHIS2. In order to do this a format of data representation has to be defined in the excel sheets. Then these excel sheets are converted to Java objects. There is a Java library, jxls which can be used to manipulate the excel files. Then the DHIS2 makes use of its importers and converters to extract the data from the Java objects and transfer them to its database.

This thesis explored the application of data mining to improve health surveillance. The survey of applications of data mining in the public health domain only highlights what is being practised currently and what challenges the Health Service institutions are facing. The health service institutions and other health care agencies can examine these applications to understand how to extract knowledge from their database systems.

For example, health institutions can arrange with government-operated medical institutions, and the body responsible for statistics in the country to gather and evaluate public health indicators. They could apply the data mining techniques in previous chapters of this thesis to locate patterns in disease outbreaks and epidemics, for example polio, per region and per district.

Health institutions can find out hitherto hidden patterns in diseases and mortality that can lead to formation of better health policies. Policies like improved vaccination planning, identification of disease vectors like malaria, cholera and how to prevent them will be better informed.

Before data mining, however, an organization must formulate clear policies on the privacy and security of patient records. Privacy is one of the issues with data mining considered in this thesis and the health institution must make sure every stakeholder, branch and district must abide by this policy.

The need for health organizations to apply data mining methods is fuelled by health fears; health concerns like sudden outbreak of epidemics, the need to predict and detect the beginning of a disease in a subtle, simple way, and the need to be more responsive to citizens.

Works Cited

- Anjaneyulu, K. S. (1998, March). Expert Systems: An Introduction. *Resonance* , pp. 46-58.
- Begg, C. B., & Berlin, J. A. (1988). Publication Bias: A Problem in Interpreting Medical Data. *Journal of the Royal Statistical Society, Series A (Statistics in Society)* , 419-463.
- Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American* , 35-43.
- Bollegala, D., Matsuo, Y., & Ishizuka, M. (2011). A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 23 (7), 1-14.
- Brooks, C., & McCalla, G. (2006). Towards Flexible Learning Object Metadata. *Int'l J. Continuing Engineering and Lifelong Learning* , 50-63.
- Heckerman, D., & Shortliffe, E. (1992). From Certainty Factors to Belief Networks. *Artificial Intelligence in Medicine*, 4 (1), 35-52.
- Chaudhri, A., & Osmon, P. (1997). Databases for a New Generation . *Object Expert* , 33-38.
- Date, C. J. (1990). *An Introduction to Database Systems*. Massachusetts: Addison-Wesley Publishing Company.
- Garcia-Abreu, A., Halperin, W., & Danel, I. (2002). *Public Health Surveillance Toolkit*. World Bank.
- Heijst, G. v., Schreiber, A., & B.J. Wielinga. (1997). Using Explicit Ontologies in KBS Development. *Int'l J. Human-Computer Studies* , 183-292.
- Hopgood, A. (1991). *Knowledge-Based Systems in Engineering*. Times Mirror Books.
- Jick, H (1977). *Adverse Effects in Relation to Renal Function*. Am J Med, 514-517
- Kaiser, G. E. (1988). Database Support for Knowledge-Based Engineering Environments. *IEEE Expert* , 18-32.
- Laffey, T. J., Cox, P. A., & Schmidt, J. L. (1988). Real-Time Knowledge-Based Systems. *AI Magazine* , 27-45.
- Leinweber, D. (1987). Expert Systems in Space. *IEEE Expert* 2(1) , 26-36.

Marsh, J., & Greenwood, J. (1986). Real-Time AI: Software Architecture Issues. *IEEE 1986 National Aerospace and Electronics Conference* (pp. 67-77). Washington D.C: IEEE Computer Society.

O'Leary, D. (1998). *IEEE* , 34-39.

Oracle (2008). *Classification. Oracle Data Mining Concepts*

Quynh, L. (2006). Issues on Health Data Collections. *Creative Dissent: Constructive Solutions* .

Rumble, J., & Smith, F. J. (1990). *Database Systems in Science and Engineering*. Briston: Adam Hilger Publication Company.

Skuce, D., & Meyer, I. (1990). Concept Analysis and Terminology: A Knowledge Based Approach to Documentation. *13th Intl. Conf. on Computational Linguistics*. Helsinki.

Weisbin, C. R. (1987). Real-Time Control: A Significant Test of AI Technologies. *IEEE Expert* 2(4) , 16-17.

Winston, P. H. (1984). *Artificial Intelligence*. Reading: Addison-Wesley.

Winter, M. (2004). Using Semantic Web Methods for Distributed Learner Modelling. *Proc. 2nd Int'l Workshop Applications of Semantic Web Technologies For E-Learning* , 236.

Wright, M., Green, M., Fiegl, G., & Cross, P. (1986). An Expert System for Real-Time Contro. *IEEE Software* , 16-24.

Wright, P. K., & Bourne, D. A. (1988). *Manufacturing Intelligence*. Massachusetts: Addison-Wesley Publishing Company.

